# Sentiment Analysis using Optimal Transport loss function

Jelena Lazić, Aleksandra Krstić and Sanja Vujnović

***Abstract*— Social networks have become an integral part of modern society, allowing users to express their thoughts, opinions, and feelings, and engage in discussions on various topics. The vast amount of user-generated content on these platforms provides a valuable source of data for sentiment analysis (SA), which is the computational analysis of opinions and sentiments expressed in text. However, most existing deep learning models for SA rely on minimizing the cross-entropy loss, which does not incorporate any knowledge of the sentiment of labels themselves. To address this limitation, we proposed a novel approach that utilizes an optimal transport-based loss function to improve sentiment analysis performance. Optimal transport (OT) metrics are fundamental theoretical properties for histogram comparison, and the proposed loss function uses the cost of the OT plan between ground truth and outputs of the classifier. The experimental results demonstrate that this approach can significantly reduce miss detections between positive and negative classes and suggest that using an OT-based loss function can effectively overcome the deficiency of existing SA models and improve their performance in real-world applications.**

***Index Terms*—sentiment analysis; natural language processing; optimal transport.**

## I. Introduction

Social networks have become an indivisible part of our lives. The number of users on Reddit, Facebook, Twitter, Instagram, Snapchat, TikTok, etc., is on the increase every day. With the increase in the number of users, there is an expansion in the number of posts on these platforms, which can generate much more complex and cross-correlated information. Online social media are a valuable channel for users to express their opinions, feelings, and thoughts, share information, and discuss their views on various topics. In the era of data-driven approaches, social network posts are an important source of data that can be very useful to overcome the lack of explicit user feedback. These posts could replace traditional polls in explanation of human activity and behavior.

In natural language processing (NLP), sentiment analysis (SA) is the computational treatment of opinions, sentiments, and subjectivity of text [1]. It is an ongoing field of research in text mining. Despite many interests in this area, it remains a non-straightforward task and has many challenges. Usually, it involves analysis of data on sentiment polarity, classification into positive and negative class, but can be extended to neutral class, or even to extreme classes, extremely positive and extremely negative. Through sentiment analysis, one can gain feedback that can improve marketing, politics, financial forecasting strategies, and many others.

There are many different approaches for text SA. Models using term frequency-inverse document frequency (TF-IDF) and word embedding [2] have been applied to a series of datasets. In their work, Dang et al. reviewed the latest studies that have employed deep learning to solve SA problems [3]. They conducted a comparative study on the experimental results obtained for the different models and input features. Except from models, there are studies of the impact of data quality on sentiment classification performance [4]. In their work, Jangid et.al applied a multi-channel convolutional neural network (CNN) and a recurrent neural network (RNN) with bidirectional long short-term memory (BiLSTM) on the microblogs and headlines of the financial domain [5]. Models based on the k-nearest neighbors (KNN) and support vector machine (SVM) have been used to classify the sentiment label of tweets [6]. For the health domain, Salas-Zárate et. al applied an ontology-based, aspect-level sentiment analysis method on tweets about diabetes [7].

Most of the introduced deep learning models for sentiment analysis are learning by minimizing the cross-entropy loss. The minus of this approach is that there is no use of any knowledge of the sentiment of labels itself. Models perform well when overall accuracy is measured, but there is no optimization over the 'weight' of the miss classifications that were made. For example, intuitively, it is known that a model which classifies a positive example as neutral, even if it is not correct, is better than a model which classifies a positive sample as negative, but this knowledge is not incorporated into the existing loss function and the model cannot learn it. In this paper, overcoming this deficiency by using optimal transport-based loss function was explored.

Optimal Transport (OT) metrics are a family of fundamental theoretical properties for histogram comparison [8]. Compared to traditional metrics, which are useful for point- based comparison of data, OT metrics consider the minimum amount of work required to transform one probability into another, considering the shape and density of the distributions, rather than simply their values at each point. The advantage of OT over traditional distance metrics in NLP is that it can handle differences in text length and word order and can capture the structure of underlaying data. It has become a thriving field, involving many researchers and many trends [9]. OT metrics are applicable to a wide range of

problems, from medicine [10- 12], and seismic wave analysis [13], to traffic modeling [14], and many others.

In NLP, there are numerous approaches based on OT metrics. Kullback–Leibler (KL) and Jensen–Shannon (JS) divergences are common loss functions for models used in NLP taskt, but they do not use label properties. To address this, R. Bhardwaj et al. investigated incorporating pre-known information about classes into the learning algorithm, intending to facilitate the model's learning algorithm using OT loss, and tested their approach on text sentiment analysis and emotion recognition in conversation [15]. In their work, J.Xu et al. used OT metrics for vocabulary construction by minimizing transfer cost to find the optimal vocabulary size [16]. Text-generating models are trained to generate the next word based on the ground-truth tokens, leading to exposure bias during evaluation when using previously generated tokens. To address this, J. Li et al. optimized the model using the OT function between these two conditions [17]. D. Alvarez-Melis et al. proposed the Gromov-Wasserstein distance, a generalization of OT, to solve the alignment problem between semantic units in cross-lingual and cross-domain problems [18]. S. Pramanick et al. propose a novel multimodal learning system for sarcasm detection based on the OT [19].

This paper is motivated by the results proposed by [15]. It is organized as follows: in the second section, an overview of optimal transport definitions is given. In the third section, methods used in experiments were explained, the optimal transport formulation, data preprocessing steps, model architecture, and the training setup were introduced. In the fourth section, there are results of the experiments. In the end, there is a conclusion.

## II. Optimal Transport

The definition of optimal transport (OT) has been addressed several times in the past [9]. It was first defined in France by Napoleon's geometer G. Monge, who defined the problem limited to the continuous distribution of mass. Monge's problem involves finding an optimal way to transport a set of dirty piles from one location to another. The cost, reflecting the difficulty or expense of moving, was assigned to each possible move. The goal in Monge's problem is to find the cheapest way to move the dirty piles. OT was redefined by the Soviet mathematician L. V. Kantorovich, who extended the definition for the whole class of linear problems, leading to the development of the tools of linear programming. He stated and provided a duality theorem that would play a crucial role in the solution of OT plans. Intuitively, the definition of OT was provided by Hitchcock [20]. Given that the cost of delivering a ton of products varies depending on the factory and the city, the goal is to find the least expensive way to transport a product. OT provides a transportation plan that minimizes these costs by optimizing the distribution of the product from the factories to the cities.

OT metrics are a family of fundamental theoretical properties for histogram comparison [8]. Given the two histograms, one can use OT to measure the minimum amount of 'work' required to transform one histogram (dirty pile, $\mu(X)$) into another (dirty pile, $\mu(T(X))$), as is shown in Figure 1. The lower overall cost of the OT plan indicates more similarities between histograms. The OT plan for minimal cost transportation of discrete source signal to discrete target signal is defined as:

$$\gamma = \arg\min_{\gamma}(< \gamma, M >) \tag{1}$$

where $\gamma$ is OT plan, its element $\gamma_{ij}$ is portion of source signal at position $i$ that should be transported to a target signal at position $j$, and $M$ is matrix of transportation cost, its element $m_{ij}$ is cost of transportation of a single unit of mass from source position $i$ to target position $j$.

Despite their intuitive formulation, computation of OT metrics involves the resolution of a linear programming cost. In his work, Cuturi smoothed the classical OT problem with an entropic regularization term [21]. Nowadays, almost all the implemented functions for OT use relaxed, regularized solutions. It is defined as:

$$\gamma = \arg\min_{\gamma}\left(< \gamma, M > + reg \cdot \Omega(\gamma)\right) \tag{2}$$

where $reg$ is regularization parameter and $\Omega$ is entropy defined as:

$$\Omega(\gamma) = -\sum_{i,j}^{\gamma_{i,j}} \log(\gamma_{i,j}) \tag{3}$$

In recent research new regularized versions of OT plans have been proposed to provide cheap computational cost and make them useful in real-life problems [22].

## III. Case Study

### A. Dataset and Preprocessing

In this work we used Twitter US Airline dataset from Kaggle [23], tweet dataset containing user opinions about United States airlines. All tweets were written in English. It has 14640 samples derived into three classes, 9178 negative, 3099 neutral and 2364 positive. The data was split randomly into train and test data, with ration 80:20, stratify to sentiment label. Train data was split once again, into train and validation data, with ration 80:20, stratify to sentiment label. On each tweet in the dataset, pre-processing steps were applied, removing retweet tag, hyperlinks, symbol # from hashtags,

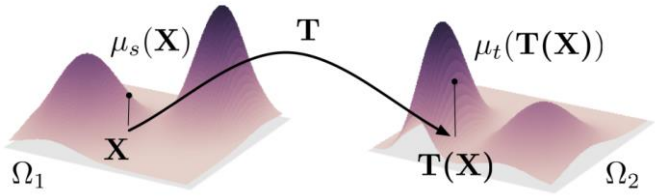


Fig. 1. OT: minimize the overall cost of moving one mass distribution onto another one. For the fixed cost matrix, more similar distributions are related to lower transport cost [8].

mentions of other users (@username) and stop words. At the end word stemming was applied.

To parse data into a suitable format, each tweet was turned into a sequence of integers, tokens, each integer being the index of a token in a dictionary. The dictionary was built from a training dataset, with a maximum length of 10000. The maximum length of sentence was set to 39, the length of 75 % of documents was less than the maximum length, and post padding was used. Sentiment labels were one-hot-encoded as positive (1,0,0), neutral (0,1,0) and negative (0,0,1).

### *B. Sentiment Analysis using Optimal Transport*

The loss function used for the SA task, based on the OT, was formulated as the cost of the OT between ground truth and the output of the model. For each document, the sentiment label was one-hot-encoded to histogram Q, and the output of the model was probability distribution P, as is shown in Figure 2. For a given sentiment Q, the loss of one document can be computed as the cost of transportation 'mass' from correct probability Q to prediction probability P. Transportation of the model's assigned masses from the embeddings of the incorrect labels to the correct label, was considered as an optimal plan. The cost matrix is given in Table 1. To smooth the loss function, entropic regularization was used with the regularization parameter 0.0001. With the lower value of the regularization parameter, results converge to non-regularized results, which tend to end in local optimal solutions (grouping all samples in two classes only).

Table 1 The cost matrix, rows represent the cost of transportation units of mass from the source signal to different positions in the target signal.

| | | model's output | | |
|---|---|---|---|---|
| | | positive | neutral | negative |
| label | positive | 0 | 0.125 | 0.25 |
| | neutral | 0.125 | 0 | 0.125 |
| | negative | 0.25 | 0.125 | 0 |

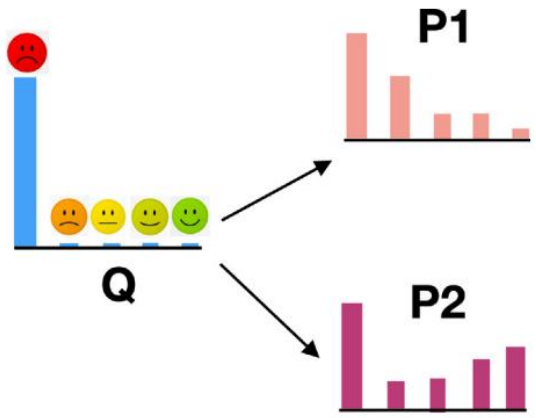


Fig. 2. P1, P2 are prediction probabilities of two sentiment classifiers and Q denotes the ground-truth distribution. Both models have the same cross-entropy, but intuitively P1 is better [15].

### *C. Neural Network Architecture*

For classification, an artificial neural network (ANN) was used. The objective of this research was not to attain the highest possible classification accuracy, but instead to compare two different loss functions. To reduce training time, less complex ANN was utilized instead of the more sophisticated models commonly used in SA tasks. The first layer was the Embedding layer, with embedding dimension 100, followed by the Global Average Pooling layer. The last two layers were fully connected: a Dense layer with 6 outputs and a relu activation function, and a Dense layer with 3 outputs and a softmax activation function. In Figure 3, the architecture of ANN is shown. Batch size was 128. Adam optimizer was used with learning rate optimized over values 0.01, 0.001 and 0.0001, where 0.001 was chosen as the best. The categorical cross-entropy loss was compared with the optimal transport loss. As OT loss cost of OT plan between ground truth and softmax outputs of the classifier was used. For both models, the number of epochs was optimized for validation of data accuracy.

## IV. Results

### *A. Categorical cross-entropy loss*

Using categorical cross-entropy as a loss function, the classifier achieved 74.85% classification accuracy. The optimal number of epochs was 15. Time spent on one epoch training was 12.4 s, averaged of 100 epochs.

Table 2. Confusion matrix of classification of test with categorical cross-entropy loss, accuracy 74.85%, the rows represent the actual or true classes, while the columns represent the predicted classes.

| | positive | neutral | negative |
|---|---|---|---|
| positive | 303 | 76 | 96 |
| neutral | 63 | 325 | 229 |
| negative | 73 | 201 | 1560 |

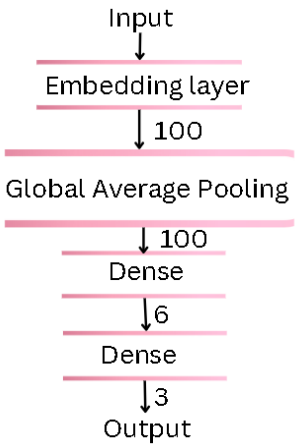


Fig. 3 ANN architecture

Table 3. Classification report for classification of test data with categorical cross-entropy loss

| | precision | recall | F1-score |
|---|---|---|---|
| positive | 0.69 | 0.64 | 0.67 |
| neutral | 0.54 | 0.53 | 0.53 |
| negative | 0.83 | 0.85 | 0.84 |

For the cross-entropy loss-based model, there were a relatively large number of miss detections between neutral and negative classes. Specifically, there were 229 neutral samples misclassified as negative, and 201 false positive negative samples misclassified as neutral, in the second row and third column, and third row and second column, respectively, in Table 2. These errors indicate that the model had difficulty distinguishing between neutral and negative sentiment, which could be influenced by imbalanced classes, negative class is much larger than neutral and positive class.

### *B. Optimal transport loss*

Using the Optimal Transport cost as a loss function, classifier achieved 76.57 % classification accuracy. The optimal number of epochs was 20. Time spent on one epoch training was 12.6 s, averaged of 100 epochs.

Table 4. Confusion matrix of classification of test with optimal transport, accuracy 76.57 %, the rows represent the actual or true classes, while the columns represent the predicted classes.

| | positive | neutral | negative |
|---|---|---|---|
| positive | 308 | 100 | 64 |
| neutral | 56 | 363 | 198 |
| negative | 57 | 210 | 1567 |

Table 5. Classification report for classification of test data with optimal transport

| | precision | recall | F1-score |
|---|---|---|---|
| positive | 0.73 | 0.65 | 0.69 |
| neutral | 0.54 | 0.59 | 0.56 |
| negative | 0.86 | 0.85 | 0.86 |

For the OT loss-based model, there were still 198 neutral samples misclassified as negative, and 210 negative samples misclassified as neutral, as is shown in Table 4. This suggests that the model still had some difficulty distinguishing between neutral and negative sentiment.

Although, there was no significant improvement in accuracy in the OT loss-based model, there was significant decrease in the number of opposite miss classifications. Using the cross-entropy loss-based model there were 96 miss detections of positive samples as negative and 73 miss detections of negative samples as positive, while using the OT loss-based model, there were 64 miss detections of positive samples as negative and 57 miss detections of negative samples as positive. Overall, it could be concluded that the OT loss-based model had lower miss detections for opposite classes (positive-negative) compared to the cross-entropy loss-based model, indicating better performance in this aspect of sentiment analysis. There was no significant difference in time spent on training or prediction with baseline and OT loss-based model.

## V. Conclusion

In this paper, sentiment analysis using optimal transport loss was researched. Traditional models for sentiment analysis are optimized over input features, data pre-processing quality, or hyper-parameter tuning. However, once these parameters are fixed, the performance of the model is limited by its loss function. Current loss functions often fail to consider the pre-existing knowledge of the labels itself, resulting in the opposite classes miss classifications, classification of positive samples as negative rather than neutral, and the reverse. To address this problem, the OT loss function was examined. Although the model's accuracy did not show a significant improvement compared to the cross-entropy loss-based model, the number of miss classifications between opposite classes (positive-negative) was substantially reduced.

Social media platforms are a primary venue for users to express their thoughts, emotions, and viewpoints, making them a valuable source of feedback. Therefore, it is crucial to enhance and refine current tools and algorithms for sentiment analysis (SA) of social media posts. In our future research, we intend to expand our investigation to address more general issues and incorporate additional classifiers for SA based on the OT loss. Additionally, we will explore other techniques for the implementation of OT algorithms.

## Acknowledgment

This research was supported by the Ministry of Science, Technological Development and Innovations of the Republic of Serbia.

## References

[1] Walaa Medhat, Ahmed Hassan and Hoda Korashz (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal volume 5, isssue 4, pages 1093- 1113

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean (2013). Distributed representations of words and phrases and their compositionality. arXiv:1310.4546

[3] Nhan Cach Dang, Marıa N. Moreno Garcıa and Fernando De la Prieta (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics 9, no. 3: 483.

[4] Lin Li, Tiong-Thye Goh and Dawei Jin (2020). How textual quality of online reviews affect classification performance: A case of deep learning sentiment analysis. Springer Neural Computing and Applications 32(3)

[5] Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah and Roger Zimmermann (2018). Aspect-Based Financial Sentiment Analysis using Deep Learning. In Proceedings of the Companion of the The Web Conference 2018 on The Web Conference, 23–27 pp. 1961–1966

[6] Mohammad Rezwanul Huq, Ahmad Ali and Anika Rahman (2017). Sentiment Analysis on Twitter Data using KNN and SVM. International Journal of Advanced Computer Science and Applications, volume 8, issue 6, pages 19-25

[7] Maria del Pilar Salas-Zárate, Jose Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Migyel Angel Rodriguez-Garcia, and Rafael Valencia-García, (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach. Computational and Mathematical Methods in Medicine

[8] Marco Cuturi and Gabriel Peyré (2018). Computational Optimal Transport. arXiv:1803.00567

[9] Cedric Villani (2008). Optimal transport: old and new. Springer

[10] Nikolay Shvetsov, Nazar Buzun, and Dmitry V. Dylov (2022). Unsupervised non-parametric change point detection in quasi-periodic signals. ACM, 32nd International Conference on Scientific and Statistical Database Management

[11] Konstantionos C. Siontis MD and Paul A. Friedman MD (2021). The Role of Artificial Intelligence in Arrhythmia Monitoring, Elseiver Card Electrophysiol Clinic, volume 13, issue 3, p543-554

[12] Jielin Qiu, Jiacheng Zhu, Michael Rosenberg, Emerson Liu and Ding Zhao (2022). Optimal Transport based Data Augmentation for Heart Disease Diagnosis and Prediction. arXiv:2202.00567

[13] Björn Engquist and Yunan Yang (2018). Seismic inversion and the data normalization for optimal transport. arXiv:1810.08686

[14] Abdullahi Adinoyi Ibrahim, Alessandro Lonardi and Caterina De Bacco (2022). Optimal transport in multilayer networks for traffic flow optimization. Algorithms, arXiv:2106.07202

[15] Rishabh Bhardwaj, Tushar Vaidya and Soujanya Poria (2019). Towards solving NLP tacks with Optimal Transport. Journal of King Saud University - Computer and Information Sciences volume 34, issue 10, part B, pages 10434-10443

[16] Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. (2021). Vocabulary Learning via Optimal Transport for Neural Machine Translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7361–7373, Online. Association for Computational Linguistics.

[17] Jianqiao Li, Chunyuan Li, Guoyin Wang, Hao Fu, Yuhchen Lin, Liqun Chen, Yizhe Zhang, Chenyang Tao, Ruiyi Zhang, Wenlin Wang, Dinghan Shen, Qian Yang, and Lawrence Carin (2020). Improving Text Generation with Student-Forcing Optimal Transport. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9144–9156, Online. Association for Computational Linguistics.

[18] David Alvarez-Melis and Tommi Jaakkola. (2018). Gromov-Wasserstein Alignment of Word Embedding Spaces. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.

[19] Shraman Pramanick, Aniket Roy, Vishal M. Patel (2022). Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 546-556.

[20] Frank L. Hitchcock (1941). The distribution of a product from several sources to numerous localities. Journal of Mathematics and Physics, 20, 224-230.

[21] Marco Cuturi (2013). Sinkhorn distances: Lightspeed Coputation of Optimal Transport distances. arXiv:1306.0895

[22] Benamou Jean-David, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré(2015). Iterative Bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing volume 37, issue 2 10.1137/141000439

[23] Kaggle: Twitter US Airline Sentiment https://www.kaggle.com/datasets/crowdflower/twitter- , Accessed 20th January 2023
7